DOCUMENT RESUME

ED 386 480 TM 024 034

AUTHOR Martinez, Michael E.; Katz, Irvin R.
TITLE Cognitive Processing Requirements of Constructed Figural Response and Multiple-Choice Items in Architecture Assessment.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY National Council of Architectural Registration Boards.
REPORT NO ETS-RR-92-5
PUB DATE Jan 92
NOTE 36p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Architecture; *Cognitive Processes; Constructed Response; *Construct Validity; Difficulty Level; *Educational Assessment; *Multiple Choice Tests; Psychometrics; Test Format; *Test Items; Test Use; Visual Stimuli
IDENTIFIERS *Figural Response Items

ABSTRACT

Contrasts between constructed response items and stem-equivalent multiple-choice counterparts typically have involved averaging item characteristics, and this aggregation has masked differences in statistical properties at the item level. Moreover, even aggregated format differences have not been explained in terms of differential cognitive processing demands of the items. In this paper, item-level differences between figural response items and their multiple-choice counterparts in architecture are examined. The figural response item format is an assessment form that uses figural materials (such as graphs, illustrations, and diagrams) as item stimuli and the medium through which knowledge and skill are demonstrated. Item-level format differences in difficulty are examined, and then whether there are corresponding differences in the cognitive processing requirements of the items that can account for the psychometric differences is studied. Based on the evidence uncovered for these connections, it is proposed that differences in processing requirements and concomitant psychometric properties might be systematic and predictable. These analyses shed light on aspects of construct validity that are frequently neglected, and they touch the interface of the usually segregated psychometric and cognitive methodologies. Four tables and three figures illustrate the discussion. (Contains 20 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-92-5

ED 386 480

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

COGNITIVE PROCESSING REQUIREMENTS OF CONSTRUCTED FIGURAL RESPONSE AND MULTIPLE-CHOICE ITEMS IN ARCHITECTURE ASSESSMENT

Michael E. Martinez
Irvin R. Katz

BEST COPY AVAILABLE

Educational Testing Service
Princeton, New Jersey
January 1992

TM024034

Cognitive Processing Requirements of Constructed Figural Response and Multiple-Choice Items in Architecture Assessment

Michael E. Martinez and Irvin R. Katz
Educational Testing Service, Princeton, NJ 08541

Running Head: ARCHITECTURE ASSESSMENT

This research was supported in part by the National Council of Architectural Registration Boards (NCARB). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of NCARB.

Copyright © 1992. Educational Testing Service. All rights reserved.

Abstract

Contrasts between constructed response items and stem-equivalent multiple-choice counterparts have yielded only a few weak generalizations. Such comparisons typically have involved averaging item characteristics, and this aggregation has masked differences in statistical properties at the item level. Moreover, even aggregated format differences have not been explained in terms of differential cognitive processing demands of the items. In this paper, we examine item-level differences between figural response items and their multiple-choice counterparts in the domain of architecture. The figural response item format is an assessment form that uses figural material (such as, graphs, illustrations, and diagrams) as item stimuli and the medium through which knowledge and skill are demonstrated. We first examine item-level format differences in difficulty and then ask whether there are corresponding differences in the cognitive processing requirements of the items that can account for the psychometric differences. After finding evidence for these connections, we propose that differences in processing requirements and concomitant psychometric properties might be systematic and predictable. These analyses are important in a larger sense in that they shed light on aspects of construct validity that are frequently neglected and they touch the interface of the usually segregated psychometric and cognitive methodologies.

Cognitive Processing Requirements of Constructed Figural Response and Multiple-Choice Items in Architecture Assessment

Comparisons of multiple-choice and constructed-response items have typically focused on data aggregated over items. The general result has been unimpressive statistical differences between the two formats. In a review of the literature, Traub and MacRury (1990) conclude that constructed response items typically are more difficult and reliable than multiple-choice counterparts, even though these differences are neither large nor consistent over studies. Important differences at the level of item statistics might be found, however. For example, in some cases the multiple-choice form of an item might be found to be more difficult than a parallel constructed-response item; this seems to be especially likely when the constructed response version is easy to answer (Martinez, 1991); in this case, the presence of the incorrect response options might be distractors in a very literal and instrumental sense. Such a hypothesis might be made more general: Format differences at the item level might be predicted and explained by the particular cognitive processing requirements of the items. This is the issue we focus on, using a constructed response item type that we refer to as figural response. The figural response item format is defined by two essential characteristics: (a) figural response items call for constructed responses—answers mentally composed by an examinee rather than chosen from a set of options (Cronbach, 1984), and (b) the response medium is figural (i.e., consisting of illustrations, graphs, etc.), rather than verbal, numeric, or some other representational form. Examinees demonstrate knowledge and skill by carrying out some operation on a figure, such as by drawing a line or arrow, or by rearranging given elements.

Figure 1 about here

The figural response items in this study assess proficiency in architecture. Figure 1 shows an item as it is presented on a computer screen. A site for a recreation center is

surrounded by structures (tennis courts, pool, etc.) which an examinee arranges on the site. Using a mouse, the examinee selects a tool button (move or rotate) on the left side of the screen. Through a combination of mouse movements and clicks, the structures are arranged in a way that satisfies the task description given in the item's stem, shown at the top of the screen. In other items, examinees respond by drawing lines or arrows, or by attaching labels to components of a diagram.

Method

Subjects

Participants were of three groups: (a) practicing architects (n= 33), (b) architecture interns (n=34), and (c) architecture students (n= 53). Practicing architects were all well-established in their profession; many were partners in their firms or professors of architecture. The interns had received a degree in architecture and had been in-training at architecture firms for two years or more, but had not yet passed the registration examinations required for professional licensure. The students ranged from first-semester college freshmen architecture students to beginning master's degree students.

Design

Figural response items were contrasted with stem-equivalent multiple-choice counterparts in a faceted test (Guttman, 1969; Snow & Lohman, 1989; Snow & Peterson, 1985). A faceted test is essentially a within-subjects experiment in which each subject answers some questions in one format and some in another. Stem-equivalence means that the task presented at the beginning of the question (the verbal stem) is constant, while the response varies, in this case between construction and selection (Traub & MacRury, 1990). In this study, multiple-choice items presented four images as response options, one of which was identified as the key.

The value of faceted tests in research is that format effects can be isolated, in large part, from test content and from person characteristics. Item format contrasts in faceted tests have potential pitfalls of carry-over effects and order effects. Carry-over effects (or

retention effects) are found when the answer to one item influences the response to a parallel item in another format; order effects are a potential confound between format and the order in which the format is given. The faceted test design used in this study (adapted from Oosterhof and Coats, 1984) attenuates these confounding effects. Parallel items were created in two formats, figural response (FR) and multiple-choice (MC). Pilot testing of FR items and an analysis of logical errors became the basis for constructing the response options for the stem-equivalent MC versions. Items were reviewed and revised by professional architecture examiners before data were collected.

Examinees answered each item in one format only. There were twenty-four items in all, comprising two test blocks of 12 items each, referred to as Block A and Block B. Half of the subjects responded to the Block A items in the FR format and to the Block B items in the MC format. The balance of the subjects responded to Block A items in the MC format and to Block B items in the FR format. By making Blocks A and B comparable, but not identical, in content, confounds between format and content were minimized for individual examinees, since similar content was sampled for both MC and FR item formats. Each item was answered in both formats, but by different groups of subjects, so the data set as a whole is free of bias from carry-over effects. Order effects were minimized by counterbalancing the order of figural response and multiple-choice items.

MC and FR items were administered via a computer-based test delivery system developed in our laboratory (Jenkins & Martinez, 1990). The system operates on an IBM-compatible microcomputer, and for the current experiment, input was entirely mouse-driven (i.e., no keyboard input was necessary). Screen images were presented on color VGA monitors offering high-resolution (640 x 480) graphics.

Materials

Twenty-four figural response items were created by architects to reflect a range of professional content, including site design, building design, and structural technologies. The items were classified into four broad categories defined by the type of knowledge

being elicited by the item.[1] Two item pairs, each pair consisting of one FR item and its MC counterpart, required the examinee to draw (for FR) or identify (for MC) an architectural symbol. We will refer to these as declarative items—items that test whether or not an examinee knows a particular "fact." Four item pairs require the examinee to apply an algorithm for solving the item, and will be referred to as learned procedure items. One such item has an examinee draw or identify a cross-section of a given topographic map. Finally, two item pairs are puzzle-like, requiring the examinee to discover a correct solution method; these items will be referred to as discovered strategy items. The site design problem (Figure 1) is a discovered strategy item: The examinee cannot simply apply a known solution method or demonstrate the possession of a single fact. Three remaining item pairs did not fit squarely into any of the above categories.

Procedure

Subjects were administered the computer-administered test in groups of six, with one computer per subject. Before taking the test, subjects were given a brief verbal introduction to the screen layout, navigation among items, and the use of on-screen tools. Subjects were given as much time as they needed in all phases of data collection. The faceted test session lasted about one hour.

Verbal protocol collection. Four of the subjects provided a concurrent verbal protocol (c.f. Ericsson & Simon, 1984) as they solved all 24 items. The verbal protocols, which involved asking subjects to "talk aloud" while solving each item, were collected to shed light on the cognitive processing requirements for figural response items and their multiple-choice counterparts.

Collection of the verbal protocols followed standard "talk aloud" procedures (Ericsson & Simon, 1984). In contrast to an interviewing technique in which subjects are

---

[1] The task analysis that led to these categories was conducted as part of a separate research effort to specify the characteristics of these architectural figural response items for diagnosis (Martinez, Katz, Sheehan, & Tatsuoka, in preparation).

directly probed for information regarding their problem-solving process, subjects were asked to say aloud freely anything that they thought. Subjects were asked not to explain their problem-solving process to the experimenter, but rather to say aloud anything that they would normally "say" to themselves during problem-solving. It is then the task of the researcher to analyze subjects' verbalizations to uncover the problem-solving strategies that would lead to those verbalizations. One subject did not follow the instructions to verbalize, so this subject's data were not included in the analysis of protocols. All four subjects' item scores were included in the psychometric analyses.

Results

Each subject was assigned two scores, one for each format, FR and MC. Two FR items were found to have extremely low p-values (.00 and .02); these and their MC counterparts were dropped from subsequent analyses. Thus, the maximum attainable score was 11 on each of the FR and MC levels of the test. A form x order (2 x 2) ANOVA using format scores as criteria indicated that neither factor had a statistically significant ($p<.05$) effect on scores. Subsequent analyses collapsed form and order.

Statistical Analysis

Figural response items were scored by two graders who were experienced in architecture assessment; one grader was a registered architect. Graders scored the items independently using a pre-established scoring rubric and later jointly resolved discrepancies in scoring. Reliability figures (Cohen's Kappa) from the scorings of the figural response problems were a mean of 0.87, and a range of 0.58 to 1. Simple ANOVAs showed that differences in mean scores between the status groups (architect, intern, and student) were statistically significant ($p<.0001$) for both FR and MC items (Table 1). The order of means followed expectations, with practicing architects having the highest scores, followed by interns and then students.

Table 1 about here

Discrimination. Item statistics for difficulty (p-values) and discriminations (r.-bis.) are shown in Table 2. Discrimination values are described first; difficulty differences across formats are then elaborated in some detail. Because each subject took items in two formats, item discrimination values could be generated for total scores of the same item type and of a different type. Item/total correlations, with the item score not removed, averaged 0.37 for figural response and 0.40 for multiple-choice. Discrimination was substantially lower when items of one format used another format as the criterion. Mean item/total correlations for figural response items predicting a multiple-choice total was a low 0.12. The value for multiple-choice items predicting a figural response total was a comparable 0.14.

Table 2 about here

Difficulty. Overall, the constructed response questions were more difficult. The mean p-value of the multiple-choice items was 0.68, compared with an average p-value for figural response items of 0.50. In nearly all of the FR/MC item pairs, the figural response version was more difficult. When items are separated by solution strategy (Table 3), differences in difficulty are not significant among the multiple-choice questions, but are significant for figural response questions. In particular, the discovered strategy questions are more difficult than are the learned procedures items.

Difference scores (MC p-value minus FR p-value) were calculated for each item and averaged within item categories (declarative, learned algorithm, and discovered strategy). The declarative and discovered strategy items show larger format differences (.28 and .37, respectively), while the learned algorithm items show a smaller format difference (.11).

Pairwise comparisons indicate that the discovered strategy/learned procedure difference is statistically significant. Small MC/FR differences for the learned procedure items cannot be attributed to ceiling effects on the MC items. Multiple-choice versions of learned procedure items typically have p-values around 0.60 or 0.70; those items with higher p-values, (e.g., SURVEY2, 0.95; VECTOR2, 0.92; see Table 2) have among the higher FR/MC differences in the learned procedure set and do not contribute to the effect. Possible mechanisms for these differences are suggested through the analysis of verbal protocols.

Table 3 about here

Protocol Analysis

In the statistical analyses, we examined item characteristics that differed between the two versions (FR and MC). The FR version of an item, for example, tended to be more difficult, but the strength of this effect seemed to be moderated by the processing demands of the particular item. Through analysis of the protocols, we sought explanations for these findings by examining in greater detail the processing requirements of the items and the methods used by subjects in solving each item.

Processing differences. Our data suggest that classification based on the processing requirements of the items might lead to predictions of format differences in item difficulty. The declarative and discovered strategy items both showed larger format differences, whereas the learned procedure items showed smaller differences. The effect found for declarative items might be explained in terms of the traditional "recognition vs. recall" distinction. MC items are usually associated with recognition; FR items with recall. This effect seems consistent with a finding by Ward, Dupree, & Carlson (1987) that simple items (in terms of number of processing steps) show stronger format differences than more complex items.

However, format effects found for the more complex items in the current study (discovered strategy vs. learned procedure) suggest that whether a particular item shows large or small format differences in difficulty has little to do with the item's complexity (number of steps) of response, contradicting the hypothesis advanced by Ward et al. Instead, we claim that format differences in item statistics are found when processes used to solve the multiple-choice and the figural response versions differ, even when the items' complexities are roughly equal; conversely, an overlap in the processing requirements in two versions of an item leads to smaller format differences. Two items demonstrate this point particularly well: the topographic map (learned procedure) and brace (discovered strategy) items, both versions of which are shown in Figures 2a, b and Figures 3a, b.

Figures 2a, 2b, 3a, & 3b about here

In the FR version of the topographic map item, the examinee must draw the cross-section corresponding the section CC on the topographic map. All three subjects appeared to solve this item by performing the following actions for each line segment drawn: (1) determine the elevation of a point on the topographic map; (2) find the corresponding point on the section graph; (3) mark that point (if beginning the problem) or draw a line from the previous point on the section graph to the new point. The procedure used to solve the MC version was similar. For each section of a MC alternative: (1) determine the elevation of a point on the topographic map; (2) find the corresponding point on the section graph; (3) decide if the corresponding parts of the two maps indicate the same level; if so, continue this process, and, if not, go on to the next MC alternative. Sample protocol excerpts illustrating comparability of processing across formats are taken from one subject solving the FR version and another subject solving the MC version of the item (TOPO6A).

FR

...Here this point down here I'm gonna to, hmmm, 86, 87. I'll start at 86 and a half....Uh., at this point here it goes down to 86. Okay from there we're gonna hover, the lowest point we're gonna go is about 85 in the middle...

MC

...This is the fourth line over so it's gonna go down to 86. 86. It's gonna drop down and come back up to 86, the last line. Right about there. That one looks pretty good. It looks pretty good. I'll look at B and see, B real quick to scan it and that's starting way up at above 91 and over on the second line.

Both subjects' protocols consist primarily of references to elevation values on the topographic map and the corresponding heights along the specified cross-section. Similar processes are used to perform both the FR and MC versions of the topographic map item. A reasonable prediction would be that subjects would find the two versions of the item similar in difficulty: a subject who cannot interpret a topographic map would not be able to solve the problem regardless of format. The item statistics correspond to this prediction: the p-values for the MC/FR version of the counterpart to this item (TOPO6A) are identical (0.62). In a comparable item (TOPO6B), the p-values differed somewhat across formats (FR = .55; MC = .75), but the difference is smaller than the mean differences for the other item sets, declarative and discovered strategy.

In other questions, particularly the discovered strategy items, the skills needed to solve the FR and MC versions of the same problem may be quite different. This is perhaps clearest on the brace problem (BRACE12A; Figures 3a & 3b). On the MC version, subjects evaluated each alternative with respect to how well the bracing prevents movement, that is, they seemed to perform a kind of mental test on the structural integrity of each design that was provided. On the FR version, subjects had to generate the bracing,

but did not evaluate the sufficiency of the bracing they produced. Protocol excerpts from separate subjects illustrate the point.

MC

...you don't really have to worry about deflection too much you just have to worry about the mobile home moving sideways and you wanna prevent that from happening. You have to look at which one is the best. Well A isn't necessarily the best. It's braced pretty well but it could still move around slightly. B is a little bit better because it has crossbracing and is connected to the anchors ...

FR

...When you're bracing you usually go in at angles so maybe, maybe I'm supposed to be reinforcing these anchors here. Of course the lines are going, at kind of across the foundations, but I believe it's my best bet...

The first subject (MC version) is concerned with the stability of structures in each option. He states this goal explicitly, observes that option A allows some movement, and then comments on the stability provided by the bracing in option B. In contrast, the second subject (FR version) describes how he will construct the bracing: by drawing diagonal lines that cross over the foundations. No mention is made of imagined movement.

If each version invites different solution strategies, there is no reason to suppose that subjects should find the MC and FR versions equally difficult. In fact, few subjects (17%) solved the FR version correctly, but a much greater percentage (75%) solved the MC version correctly. The statistics are comparable for a parallel item pair (BRACE12B): FR = 30% and MC = 73%.

Strategies for using response options. The items used in this study were developed to assess a range of architectural knowledge and skill. Accordingly, a variety of methods were used to solve them: a topological map item was approached differently from a brace

item. While problem-solving on the FR versions of each item could be characterized only by a very general model of goal-directed behavior (Katz, Martinez, Sheehan, & Tatsuoka, in preparation), their problem-solving on the MC versions was more restricted. Almost all problem-solving on the MC items could be characterized in one of two ways: (a) judge each option on own merits (local evaluation), or (b) solve the problem by comparing options and eliminating less plausible ones (global evaluation).

For each of the three protocol subjects, the problem-solving strategy used on each MC item was classified as either global, local, or unknown. The strategy was categorized as "global" if there was evidence in the verbal protocol that the subject considered more than one option at a time or compared options to one another. If a subject decided on the correctness of an option without comparing it to another, and before going on to the next option, the strategy was categorized as "local". A strategy was labeled "?" if the subject's verbalizations did not provide enough information to decide between the two strategy categories.

Table 4 about here

The results are shown in Table 4. There is little regularity in whether a particular item is solved using one strategy or another; in fact, no item was solved in the same way by all three subjects. The only pattern discernable is a tendency for Subject 2 to use a local strategy and Subject 3 to use a global strategy. The data therefore show intra-subject strategic differences across items and hint at inter-subject preferences for certain strategies. Moreover, the presence of response options is non-trivial in their affect on the course of problem solution.

Discussion

According to Messick (1989, p. 17), "construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores." Ideally, this

evidence comes from many sources, including empirical findings and theoretical work, and an emerging picture of validity evolves over time. To this point, item difficulty and discrimination, aggregated over items, have typically been the basis for contrasting constructed-response and multiple-choice formats. These contrasts have revealed that, generally, constructed response formats are more difficult and discriminating than their multiple-choice counterparts (Traub & MacRury, 1990). This finding was corroborated in a previous study involving paper-and-pencil figural response items in pre-college science assessment (Martinez, 1991). Yet multiple-choice items are not always easier than their constructed-response counterparts, and differences between item pairs varies appreciably. Based on our findings, we propose that the cognitive processing requirements of particular items shed some light on why item properties across formats are sometimes large and sometimes small.

Constructed response items have usually been found to be more discriminating than multiple-choice counterparts (Traub & MacRury, 1990; Martinez, 1991). That is, constructed response items are more efficient in predicting a total score and this efficiency has usually been indexed by a pt.-biserial or r-biserial correlation between an item score and a total score of the same item format. We found little difference in prediction across the formats. Constructed response measurement might typically be more discriminating than multiple-choice, because guessing correctly is not a factor in adding irrelevant variance to the test scores. The predictive power of constructed response items might be attenuated by that fact that scoring unreliability sometimes plays a significant role. In the case of this assessment, items were drawn from diverse aspects of architecture. Perhaps a lack of homogeneity of content contributed to discrimination values that are somewhat low and that offered no advantage for figural response items as a constructed response format.

In this study, analyses of verbal protocols shed light on the cognitive processing requirements of figural response items and their multiple-choice counterparts. We found evidence that processing of stem-equivalent items might be similar or different depending

on the specific demands of item. If the salient processing demands are the application of skills, as in the topographic map items, there may be little differences in processing across formats and therefore little difference in item statistics. A different story emerged with the bracing items, in which subjects had to construct the response in the figural response version, but tended to evaluate the multiple-choice options on the basis of their implications for structural integrity. With that item pair, and with discovered strategy items generally, item statistics differed across format. While the recall versus recognition distinction may in some cases describe the format effects occurring with the declarative items, it is not a sufficient explanation of the discovered strategy format effects The distinction between construction and evaluation might be a better alternative or at least a supplemental hypothesis for explaining the differences.

The construction/evaluation dichotomy has also been drawn in at least one other domain: computer programming. McKendree & Anderson (1987) demonstrated that it was possible for subjects to be able to generate correct computer code, but be unable to evaluate similar code presented to them (i.e., be able to say what would happen when the code executed on a computer). In a related study of LISP computer programming, Kessler (1988) demonstrated that learning to create short computer programs did not aid subjects' ability to evaluate (mentally run) similar computer programs, and vice-versa.

For the brace problem, the skills associated with solving the MC version (via evaluation) and the FR version (via generation) appear to be different. Recall that differences in difficulty between formats on the topographic items was negligible, but differences were large on the bracing items. We had no a priori grounds for saying that evaluation would be easier than generation, but the very fact of large differences in p-values we find provocative and view as a possible (if not demonstrable) linkage between cognitive and psychometric realms. The generation of such linkages may be the most pressing task for a cognitive psychology of measurement.

In our protocols, we found evidence that multiple-choice response options were frequently used in a manner referred to by Snow (1980) as a response elimination strategy. In response elimination, item options are pared down as some are eliminated on the basis of implausibility, and subjects make the best choice or else guess from the remainder. With another strategy, constructive matching, an item's answer is mentally invented and the responses are searched for a match. In a kind of elaboration of Snow's model, our protocols indicated that response options can be evaluated locally and serially, or globally and more in parallel. In some cases, especially when a response is elaborate, a response construction strategy is highly unlikely because the the answer would exceed the capacity of short-term memory. Some sort of response elimination strategy is apparently unavoidable in the topographical map/section question because its is unlikely that one could hold the entire solution in working memory while simultaneously mapping between topographical and cross-sectional representations. The presence of response options invites strategizing which we argue has little to do with the target construct and which may compromise that validity of the measure. But we also conjecture that any item format, including CR formats, will be amenable to strategies that test developers probably did not intend.

Conclusions

In part, this paper illustrates the importance of multiple perspectives in elucidating the nature of the construct embodied in an assessment method. The meaning of what is measured was illuminated by the items' psychometric properties, the cognitive processes used to answer the items, and their interrelations. Other theoretical perspectives and methodologies are of course possible and necessary, including examining the relationship between item format and mental abilities. For example, one reasonable hypothesis is that ability to demonstrate knowledge within a figural medium draws from one kind or another of "figural" aptitudes, such as visualization or visual memory (Ekstrom, French, & Harman, 1976). Relationships between aptitudes and item types can be investigated using

a research methodology similar to that used here, in which a faceted test is administered along with measures of the mental abilities of interest. Format scores are then regressed onto aptitude scores and differential relationships are indicated by differences in slope between simple regression lines (Cronbach & Snow, 1977; Snow & Lohman, 1989). Of course, descriptions of relationships between assessment format and aptitudes does not say much about the cognitive underpinnings of the two. Cognitive research methods could contribute to an understanding of performance on aptitude tests, the test item format of interest, and common cognitive components.

The intent of this study was to examine cognitive and psychometric differences across item formats, but the research methodology has significance beyond illuminating the multiple-choice/constructed response distinction. This is one of a small but growing number of studies that seek to link cognitive and correlational paradigms (Cronbach, 1957). Such an analysis can be an important aspect of construct validity research (Messick, 1989), both for existing item formats in large-scale use and for understanding the possible meanings and values of new forms of assessment. Such meanings might be important even if they are not distinguished by psychometric data aggregated at the test level (Frederiksen & Collins, 1989).

References

Cronbach, L. J. (1957). The two disciplines of scientific psychology. American Psychologist, 12, 671-684.

Cronbach, L. J. (1984). Essentials of psychological testing. New York: Harper & Row.

Cronbach, L. J., & Snow, R. E. (1977). Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). Kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis. Cambridge, MA: The MIT Press.

Frederiksen, J. R. & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18 (9), 27-32.

Guttman, L. (1969). Integration of test design and analysis. Proceedings of the 1969 invitational conference on testing problems. Princeton, NJ: Educational Testing Service.

Jenkins, J., & Martinez, M. E. Figural Response Authoring and Measurement Environment. [computer program]. Princeton, NJ: Educational Testing Service.

Katz, I. R., Martinez, M. E., Sheehan, K., & Tatsuoka, K. K. (manuscript in preparation). Diagnostic assessment in architecture.

Kessler, C. M. (1988). Transfer of programming skills in Novice LISP Learners. Unpublished doctoral dissertation. Carnegie Mellon University, Pittsburgh, PA.

Martinez, M. E. (1991). A comparison of multiple-choice and constructed figural response items. Journal of Educational Measurement, 28, 131-145.

McKendree, J. M., & Anderson, J. R. (1987). Effects of practice on knowledge and the use of basic LISP. In J. Carroll (Ed.), Interfacing thought: Cognitive aspects of human-computer interaction. Cambridge, MA: The MIT Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). Educational measurement (3rd edition). New York: Macmillan.

Mislevy, R. J. (in press). A framework for studying differences between multiple-choice and free-response test items. In R. E. Bennett & W. C. Ward (Eds.), Construction vs. choice in cognitive measurement. Hillsdale, NJ: Erlbaum.

Oosterhof, A. C., & Coats, P. K. (1984). Comparison of difficulties and reliabilities of quantitative word problems in completion and multiple-choice item formats. Applied Psychological Measurement, 8, 287-294.

Snow, R. E. (1980). Aptitude processes. In R. E. Snow, P.-A. Federico, & W. E. Montague (Eds.), Aptitude, learning, and instruction, Vol. 1: Cognitive process analyses of aptitude. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.). Educational measurement (3rd edition). New York: Macmillan.

Snow, R. E., & Peterson, P. L. (1985). Cognitive analyses of tests: Implications for redesign. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics. Orlando, FL: Academic Press.

Traub, R. E., & MacRury, K. (1990). [Multiple-choice vs. free-response in the testing of scholastic achievement]. In K. Ingenkamp & R. S. Jäger (Eds.), Tests und trends 8: Jahrbuch der Pädagogischen Diagnostik. Weinheim und Basel: Beltz Verlag.

Ward, W. C., Dupree, D., & Carlson, S. B. (1987). A comparison of free-response and multiple-choice questions in the assessment of reading comprehension. Educational Testing Service Research Report (RR-87-20). Princeton, NJ: Educational Testing Service.

Table 1

ANOVA: Multiple-Choice and Figural Response Format Scores, by Status Group

	Status Groups						
	Students (n=53)		Interns (n=34)		Architects (n=33)		
Format	M	SD	M	SD	M	SD	Significance Test
Figural Response	4.83	1.82	5.59	1.64	6.70	1.51	F (2, 117) = 12.5*
Multiple-Choice	6.58	2.07	7.91	1.40	8.48	1.42	F (2, 117) = 13.7*

*differences between means are statistically significant at p<.0001.

Table 2

Item Statistics for Figural Response Items and Stem-Equivalent Multiple-Choice Counterparts

Item	p-value		r.-bis.		Solution Strategy
	FR	MC	FR	MC	
CURB17	0.08	0.47	0.14	0.59	declarative
CURB17B	0.17	0.45	0.63	0.36	declarative
WINPROJ	0.78	0.92	0.83	0.18	declarative
WINCASE	0.55	0.85	0.54	0.67	declarative
LIVELOAD10A	0.72	0.73	0.67	0.46	learned procedure
LIVELOAD10B	0.68	0.72	0.53	0.63	learned procedure
SURVEY1	0.48	0.57	0.65	0.57	learned procedure
SURVEY2	0.78	0.95	0.43	0.75	learned procedure
TOPO6A	0.62	0.62	0.34	0.60	learned procedure
TOPO6B	0.55	0.75	0.57	0.68	learned procedure
VECTOR1	0.63	0.87	0.34	0.34	learned procedure
VECTOR2	0.75	0.92	0.52	0.82	learned procedure
SITEPLAN	0.40	0.58	0.16	0.46	discovered strategy
SP2	0.48	0.80	0.58	0.66	discovered strategy
BRACE12A	0.17	0.75	0.43	0.55	discovered strategy
BRACE12B	0.30	0.73	0.47	0.67	discovered strategy
WELD 4A	0.80	0.63	0.34	0.76	mixed
WELD 4B	0.07	0.08	0.44	-0.01	mixed
ROOF19B	0.40	0.63	0.83	0.60	mixed
VAPORB	0.57	0.62	0.18	0.33	mixed
SEISJ11	0.53	0.67	0.46	0.55	mixed
SEISJ11B	0.60	0.67	0.45	0.52	mixed

Table 3

ANOVA: Mean P-Values for Multiple-Choice (MC), Figural Response (FR), and MC-FR

	Solution Strategy						
	Declarative (4 items)		Learned Procedure (8 items)		Discovered Strategy (4 items)		
Format	M	SD	M	SD	M	SD	F (2, 13)
Multiple-Choice	.68	.25	.77	.14	.71	.09	0.46, n.s.
Figural Response	.40	.33	.65	.10	.34	.13	4.93, p<.05*
MC-FR	.28	.10	.11	.09	.37	.17	7.44, p<.01*

*Scheffe procedure shows Discovered Strategy differs from Learned Procedure at p<.05.

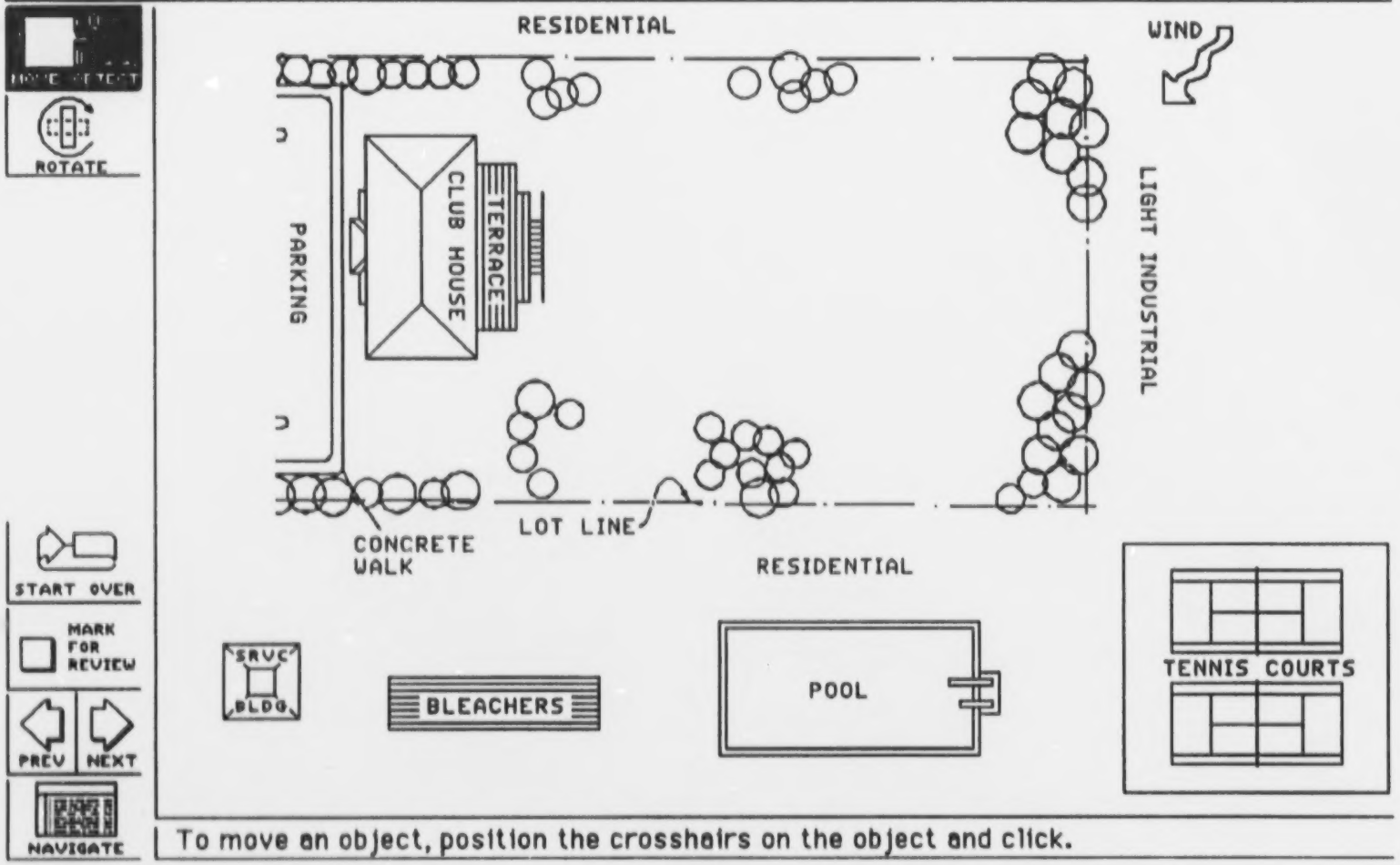
Table 4

Solution Strategies Used by Subjects

Item	Subjects		
	1	2	3
topo	local	local	global
liveload	global	local	global
window	?	local	global
site	global	global	local
survey	?	?	?
curb	global	local	global
vapor	?	local	global
seismic	?	local	?
brace	global	local	local
vector	?	global	?
weld	local	?	global
TOTALS	3 local	7 local	2 local
	4 global	3 global	6 global
	5 unknown	2 unknown	4 unknown

NCARB ncarb siteplan #3 of 24 Status: Not Attempted ID: Time -

A recreational center site plan must accommodate a club house in its present position, as well as tennis courts, pool, bleachers, and a service building. Prepare the site plan according to the following objectives: (1) Preserve all trees. (2) Bleachers shall serve the tennis courts. (3) Pool shall be adjacent to the clubhouse. (4) Service building shall relate to the club house and the parking lot.



To move an object, position the crosshairs on the object and click.

Figure 1. Sample figural response item.

NCARB ncarb topo6a #17 of 24 Status: Not Attempted ID: Time -

On the Section Grid below, draw the section that corresponds to the section CC indicated on the topographical map.

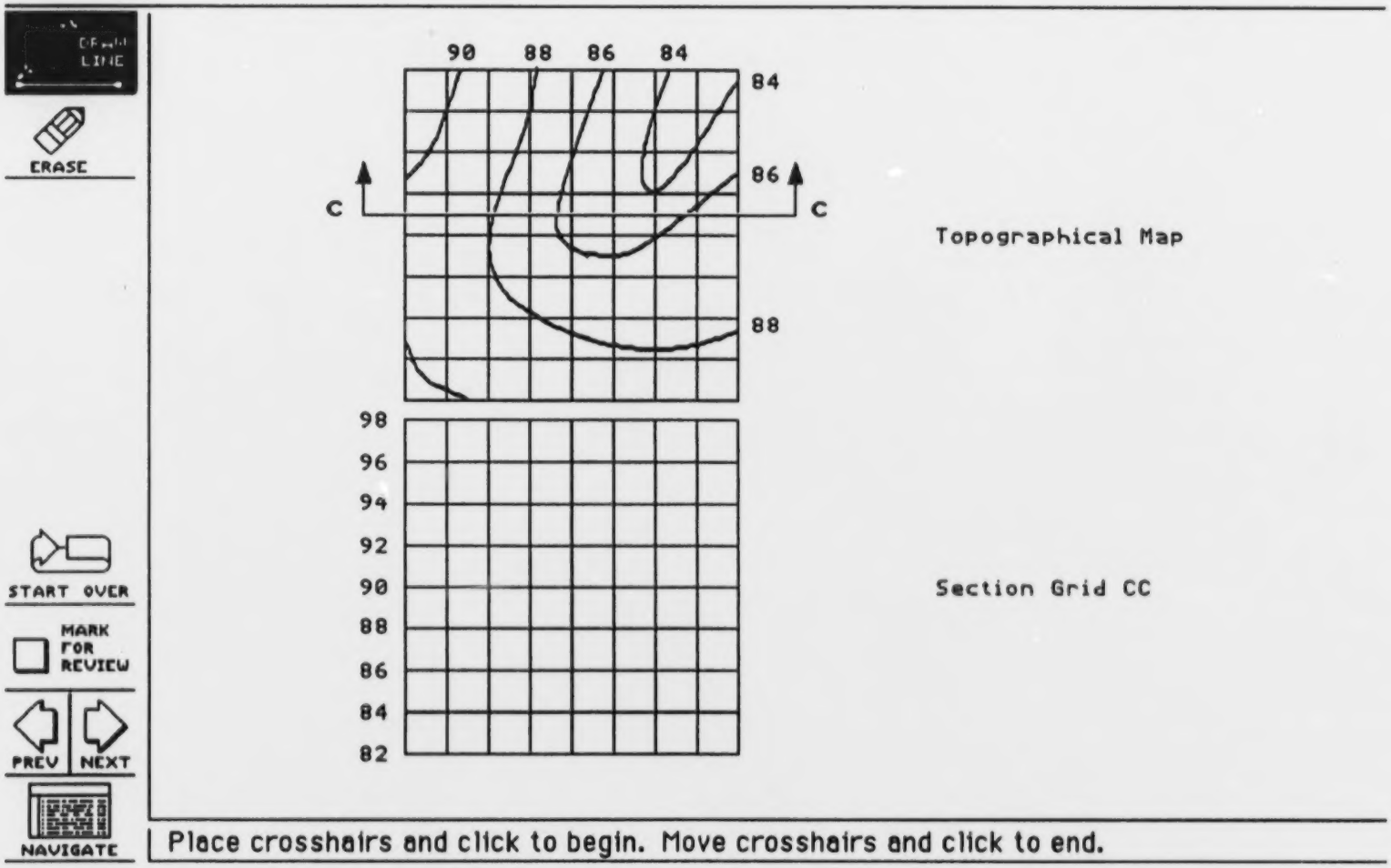
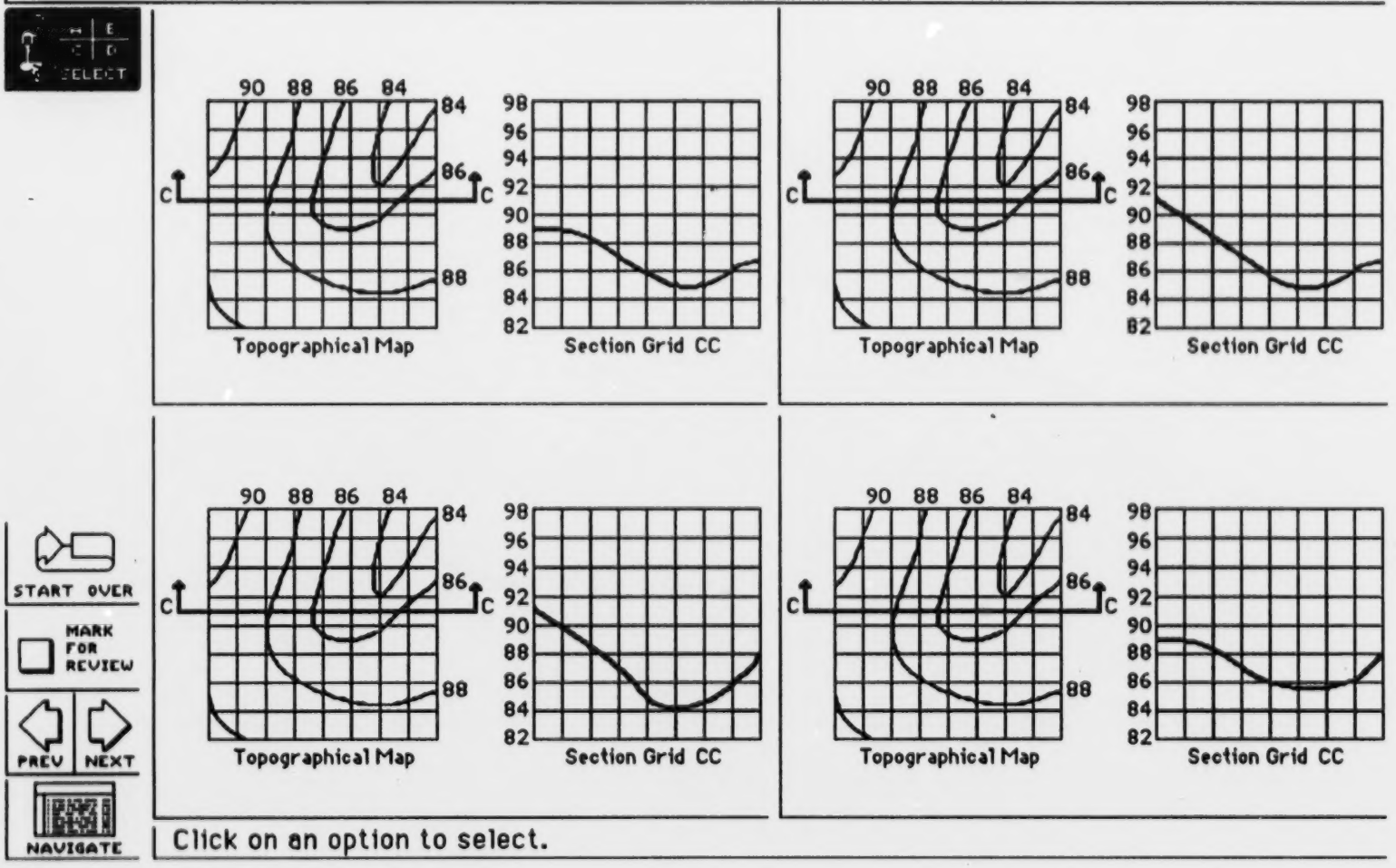


Figure 2a. Topographical map item: figural response format.

NCARB Figural Response tcc_m #14 of 24 Status: Not Attempted ID: Time -

Select the section that corresponds to section CC indicated on the topographical map.

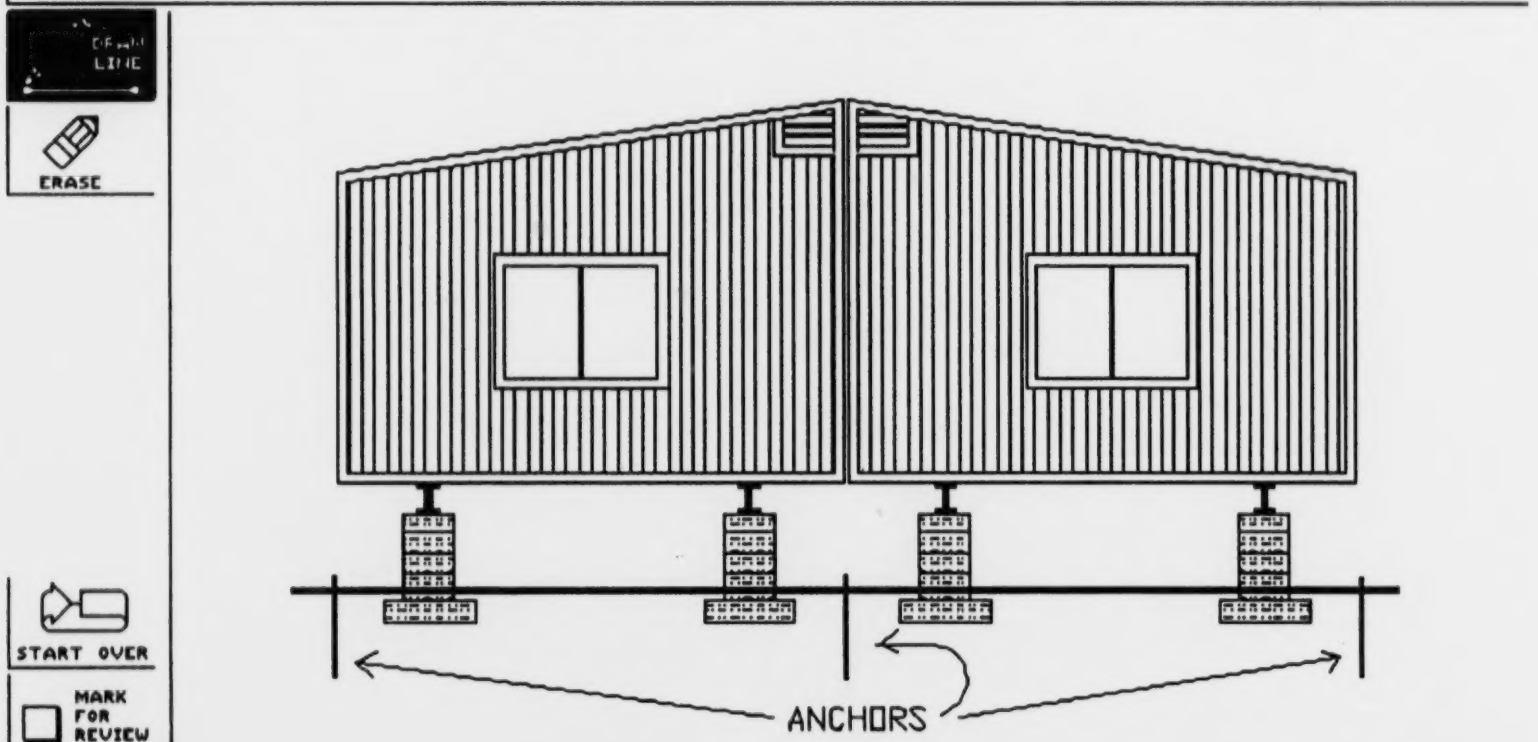


Click on an option to select.

Figure 2b. Topographical map item: multiple-choice format.

NCARB ncarb bracel2a #19 of 24 Status: Not Attempted ID: Time -

Draw the lateral force cable bracing for the foundation piers on a double wide mobile home.

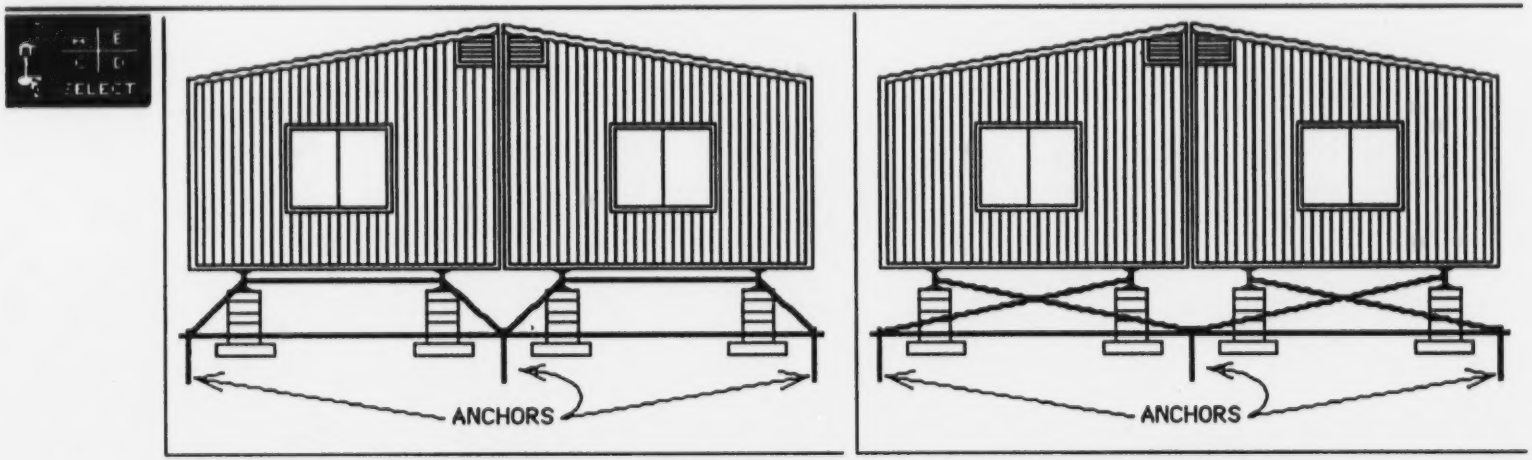


Place crosshairs and click to begin. Move crosshairs and click to end.

Figure 3a. Cable bracing item: figural response format.

NCARB Figural Response bra_m #1 of 24 Status: Not Attempted ID: Time -

Indicate the proper location for lateral force cable bracing for the foundation piers on a double-wide mobile home by selecting the proper illustration.



SELECT

START OVER

MARK FOR REVIEW

PREV NEXT

NAVIGATE

Click on an option to select.

Figure 3b. Cable bracing item: multiple-choice format.